Vorlesung 14b

Kann das Zufall sein?

Beispiele von statistischen Tests

Beispiel 1:

“Passen die Verhältnisse in den Rahmen?”

Fishers exakter Test

(vgl. Buch S. 130/131

Aus einer Urne mit 80 roten und 87 blauen Kugeln

wurden 113 Kugeln entnommen.

40 davon waren rot, und 73 waren blau.

Passt das zur Hypothese, dass die Kugeln

rein zufällig gezogen wurden?

Stimmen die Verhältnisse einigermaßen,

oder fallen sie aus dem Rahmen?

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Unter den 113 gezogenen Kugeln erwartet man das

Verhältnis 80:87 für rot zu blau

Tatsächlich ist das Verhältnis 40: 73 viel günstiger für blau!

Wie lässt sich das quantifizieren?

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Unter des Hypothese des rein zufälligen Ziehens

ist die Anzahl X der gezogenen roten Kugeln

hypergeometrisch verteilt

mit Parametern $n = 113$, $g = 167$, $r = 80$.

Dafür ergibt sich:

$$\mathbf{E}[X] = n \cdot \frac{r}{g} = 54.1 \; .$$

$$g = 167,\ r = 80,\ n = 113$$

X ist Hyp(n, g, r)-verteilt

$$\mathbf{E}[X] = n \cdot \frac{r}{g} = 54.1\ .$$

Zur Erinnerung:

$$\mathbf{P}(X = k) = \frac{\binom{r}{k}\binom{g-r}{n-k}}{\binom{g}{n}}\,, \quad k = 0, 1, \ldots, n\ .$$

	gezogen	nicht gezogen	Summe
rot	**40**	40	80
blau	73	14	87
Summe	113	54	167

Die Wahrscheinlichkeit, ein Ergebnis zu erhalten,

das mindestens so weit von 54 weg ist

wie der beobachtete Wert 40, ist

$$\mathbf{P}\big(|X - 54| \geq |40 - 54|\big)$$

$$= \mathbf{P}(X \leq 40) + \mathbf{P}(X \geq 68)$$

$$= 5.57 \cdot 10^{-6}\,.$$

$$\mathbf{P}\Big(|X - 54| \geq |40 - 54|\Big) = 5.6 \cdot 10^{-6}$$

Was bedeutet das?

Fazit: Angenommen die Hypothese trifft zu.

Dann tritt ein Ergebnis, das so extrem ist wie das beobachtete, gerade 6 mal in einer Million auf.

Damit wird die Hypothese mehr als fragwürdig.

Man nennt die berechnete Wahrscheinlichkeit

den zu den Daten gehörigen p-Wert

oder auch das *beobachtete Signifikanzniveau*,

zu dem die Hypothese abgelehnt wird.

Wie passt unser Urnen-Beispiel in die Welt?

Es geht um die Fragestellung

“Passen die Proportionen

– oder werden die Chancen beeinflusst?”

Zwei Verpackungen einer Botschaft – und eine Frage dazu:

A. Die “sanfte” Therapiemethode T1 brachte
in nicht weniger als 30% der Fälle keinen Heilungserfolg,
wohingegen die harte Therapiemethode T2
in immerhin 80 % der Fälle erfolgreich war.

B. Sogar die harte Therapiemethode T2 brachte
in nicht weniger als 20% der Fälle keinen Heilungserfolg,
wohingegen die sanfte Therapiemethode T1
in immerhin 70 % der Fälle erfolgreich war.

Welche Therapiemethode würden Sie (als Arzt) bevorzugen?

Von insgesamt 167 Ärzten

wurden rein zufällig 80 ausgewählt,

denen die Botschaft in der Form A vermittelt wurde,

die restlichen 87 bekamen die Botschaft in der Form B.

Jeder der Ärzte hatte sich daraufhin für die Bevorzugung

einer der beiden Therapiemethoden zu entscheiden.

Das Ergebnis war:

	für Methode T1	für Methode T2	Summe
A	40	40	80
B	73	14	87
Summe	113	54	167

Für das Testen der Hypothese

“Die Verpackung der Botschaft

hat keinen Einfluss auf die Entscheidung”

eignet sich das eingangs besprochene Urnenmodell.

Unter dieser Hypothese

kommt die Aufteilung der $80 + 87$ Formulare

auf die 113 Befürworter von T1

und die 54 Befürworter von T2

rein zufällig zustande.

So gesehen kann das Ergebnis “wohl kam Zufall sein”:

unter unserer Hypothese tritt ein Ausgang,

der so extrem ist wie der beobachtete,

gerade mal 6 mal in einer Million auf.

Eine weitere Möglichkeit zum Testen der Hypothese

“Zwei Verhältnisse sind gleich”

bietet die *Normalapproximation*.

Vergleiche dazu unsere Übungsaufgabe 45b.

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

Anteilsschätzung über die Normalapproximation:

X := # roter Kugeln bei $n = 113$ Zügen

X ist Hyp$(113, 167, 40)$-verteilt

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

$$H = \frac{X}{n}$$

$$\sigma_H^2 = \frac{p(1-p)}{n}\frac{g-n}{g-1} = 0.022 \cdot 0.326 = 0.000718$$

$$\sigma_H = 0.0268$$

$$Z := \frac{H - 80/167}{\sigma_H} \text{ ist approximativ N}(0,1)\text{-verteilt}$$

(trotz der schwachen Abhängigkeiten beim Ziehen ohne Zurücklegen, vgl. die beiden letzten Folien von Vorlesung 8a)

	gezogen	nicht gezogen	Summe
rot	40	40	80
blau	73	14	87
Summe	113	54	167

$$Z := \frac{H - 80/167}{\sigma_H}$$ ist approximativ N$(0, 1)$-verteilt

Der beobachtete Wert von Z war $z = -4.67$.

$$\mathbf{P}(|Z| > 4.67) = 3 \cdot 10^{-6}$$

ist hier der p-Wert, zu dem die Hypothese abgeleht wird.

Beispiel 2:

“Kann *diese* Verschiebung des Mittelwertes Zufall sein?”

Der t-Test:

n Messwerte $x_1, \ldots, x_n$ haben den Mittelwert m

(alles gemessen auf einer bestimmten Skala.)

Denken wir an $n = 16, \;\; m = \mu_0 + 0.75$.

Dabei ist μ_0 ein vorgegebener “Sollwert”. Unterscheidet sich der beobachtete Mittelwert m signifikant von μ_0?

Eine erste Auskunft gibt ein Vergleich

des Unterschiedes $|m - \mu_0|$

mit dem *Standardfehler*

$$f := s/\sqrt{n}.$$

Nehmen wir an: $s = 1.2$.

Dann ist $f = s/\sqrt{n} = 1.2/\sqrt{16} = 0.3$, und

m unterscheidet sich “um 2.5 Standardfehler” von μ_0.

Wenn wir die folgende Modellannahme treffen:

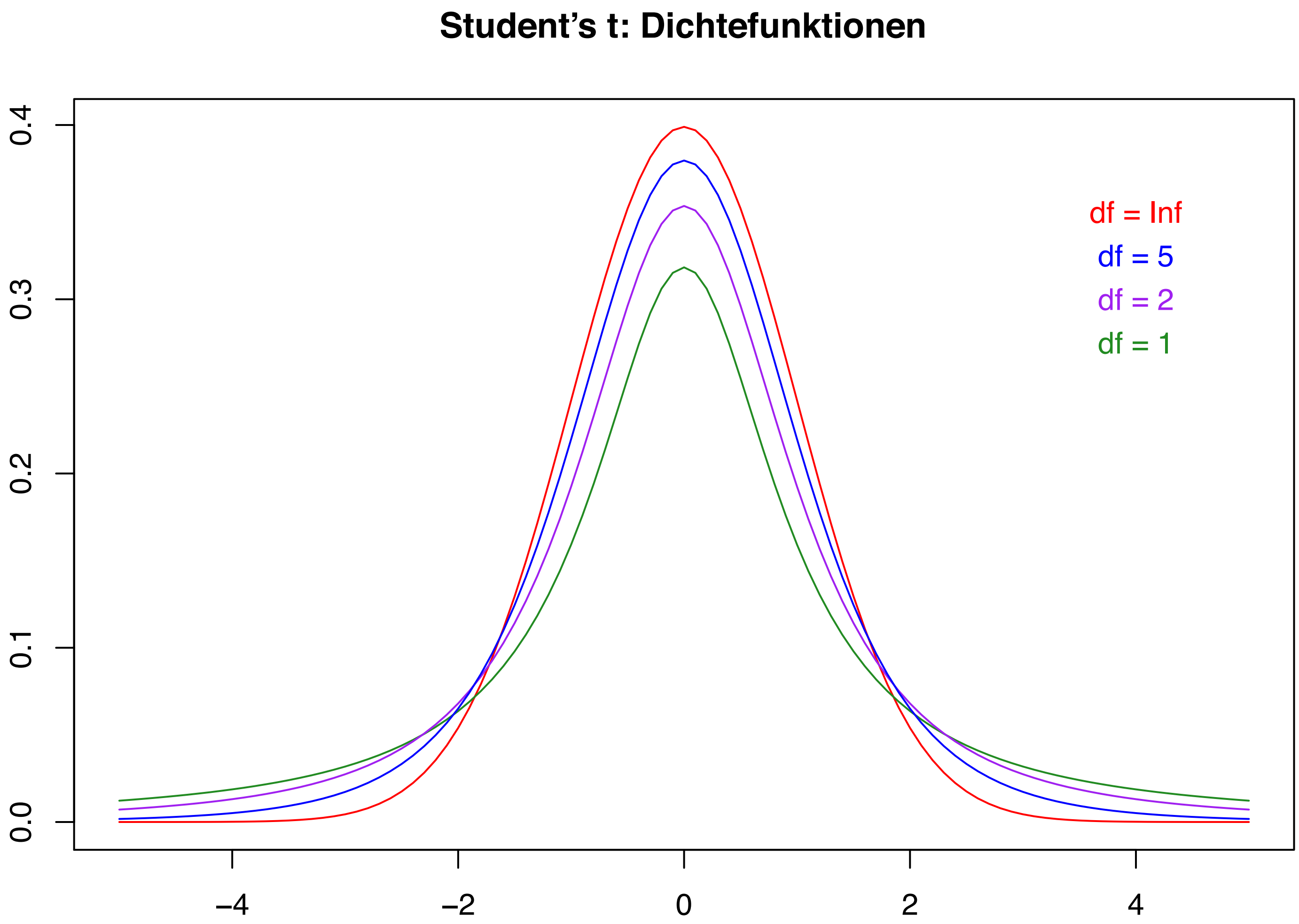
$x_1, \ldots, x_n$ sind Realisierungen von unabhängigen,
N(μ, σ^2) verteilten Zufallsvariablen,

dann können wir die Aussage
“Wie signifikant unterscheidet sich m von μ_0 ?”
sogar exakt beantworten.

Unter der Hypothese “$\mu = \mu_0$” kommt es zu einem Unterschied von $\bar{X}$ und μ_0, der mindestens so groß ist wie der beobachtete, mit der Wahrscheinlichkeit

$$\mathbf{P}(|T_{15}| \geq 2.5) = 0.025.$$

Denn dann ist $(\bar{X} - \mu_0)/(s/\sqrt{n})$ t-verteilt mit $n-1$ Freiheitsgraden (siehe Vorlesung 14a).



Dichten von T_{df}

Unter der Hypothese "$\mu = \mu_0$" kommt es zu einem Unterschied von $\bar{X}$ und μ_0, der mindestens so groß ist wie der beobachtete, mit der Wahrscheinlichkeit

$$\mathbf{P}(|T_{15}| \geq 2.5) = 0.025.$$

Man spricht vom p-Wert für die Ablehnung der Hypothese $\mu = \mu_0$ zugunsten der Alternative $\mu \neq \mu_0$.

Oft gibt man sich ein *Signifikanzniveau* α vor.

Wenn der p-Wert kleiner als α ist, sagt man: Die Hypothese $\mu = \mu_0$ kann zugunsten der Alternative $\mu \neq \mu_0$ zum Niveau α abgelehnt werden.

Populär ist die Wahl $\alpha = 0.05$:

Wenn der p-Wert kleiner als 0.05 ist, sagt man auch kurz: m ist (nach dem t-Test) signifikant von μ_0 verschieden.

Beispiel 3:

“Unterscheiden sich zwei Mittelwerte signifikant?”

Der t-Test für ungepaarte Stichproben.

Die Mittelwerte m_x und m_y

von zwei Stichproben des Umfangs n_x und n_y

unterscheiden sich um 0.5 Einheiten: $|m_y - m_x| = 0.5$.

Ist dieser Unterschied signifikant?

Nach bewährtem Rezept vergleichen wir $|m_x - m_y|$

mit “seinem Standardfehler” f.

Weil wir an unahängige Stichproben denken,

addieren sich die Varianzen:

$$f := \sqrt{f_x^2 + f_y^2}$$

Eine Maßzahl für den “relativen Unterschied” ist also

$$\frac{m_y - m_x}{f}.$$

Interpretiert man die x_i und die y_j

als Realisierungen von

unabhängigen Zufallsvariablen X_i, Y_j,

(mit (X_i) identisch verteilt, (Y_j) identisch verteilt)

dann stellt sich die Frage nach der Verteilung von

$$T := \frac{\bar{Y} - \bar{X}}{F} = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}}}$$

$$T := \frac{\bar{Y} - \bar{X}}{F} = \frac{\bar{Y} - \bar{X}}{\sqrt{\frac{S_X^2}{n_x} + \frac{S_Y^2}{n_y}}}$$

Für große n_x, n_y ist T annähernd N$(0, 1)$-verteilt

(wegen des Zentralen Grenzwertsatzes und des Gesetzes

der großen Zahlen).

Was aber ist für kleine n_x, n_y?

Hier kommt man zumindest unter der zusätzlichen Annahme

weiter, dass die X_i und Y_j normalverteilt sind.

Man kann zeigen, dass T dann annähernd t-verteilt ist
mit einer i.a. nicht ganzzahligen Anzahl von Freiheitsgraden.

Die Formel dafür (die man sich nicht merken muss) findet man auf
http://en.wikipedia.org/wiki/Student's_t-test
im Abschnitt "Unequal sample sizes, unequal variance"

Wichitger ist der praktische Umgang damit in R,
zu dem man dort auf die Frage ?t.test Auskunft bekommt.

Nimmt man überdies an (was nicht immer gerechtfertigt ist!) ,
dass die beiden Varianzen gleich sind ($\sigma_X^2 = \sigma_Y^2 =: \sigma^2$),
dann gibt es eine effizientere Art für die Schätzung von σ:

$$s_{X,Y} := \sqrt{\frac{1}{n_x+n_y-2}\Big((X_1-\bar{X})^2+\cdots+(Y_{n_y}-\bar{Y})^2\Big)}$$
$$= \sqrt{\frac{(n_x-1)s_X^2+(n_y-1)^2 s_Y^2}{n_x+n_y-2}}$$

Der Charme davon ist, dass die entsprechende t-Statistik nicht nur approximativ, sondern exakt t-verteilt ist. Mit einem “Projektionsargument à la Fisher” zeigt man wie im Satz von Gosset und Fisher:

$$\frac{(\bar{X}-\mu_X)-(\bar{Y}-\mu_Y)}{\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}\; s_{X,Y}} \quad \text{ist } \mathrm{t}(n_x+n_y-2)\text{-verteilt.}$$

Beispiel 4:

“Ist der Mittelwert (von Differenzen)

signifikant von Null verschieden?”

Der t-Test für gepaarte Stichproben.

Welches Schuhsohlenmaterial nutzt sich weniger ab? *

10 Jungen, 10 Paar Schuhsolen (je 5 vom Material A bzw. B)

2 Varianten der Versuchsplanung (eine schlechte, eine gute):

1) Wähle zufällig 5 der 10 Jungen und gebe ihnen Material A, die anderen bekommen Material B.

2) Jeder Junge erhält Material A für den einen und Material B für den anderen Fuß.

*nach einem Bsp. im Buch “Statistics for experimenters” von G.E.P. Box, W.G.Hunter, J.S. Hunter, Wiley 1978, 2005

Welches Schuhsohle (A oder B) nutzt sich weniger ab?

Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Dieses Datenmaterial kommt aus einem

gemäß Variante 2 durchgeführten Versuch.

Welches Schuhsohle (A oder B) nutzt sich weniger ab?

Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Betrachte die Differenzen $D_i := A_i - B_i$

als unabhängig und N(μ, σ^2)-verteilt.

Der t-Test mit der Hypothese $\mu = 0$

ergibt einen p-Wert von 0.008,

also einen hochsignifikanten Unterschied.

(Die Variabilität zwischen den Individuen
geht hier nicht in die t-Statistik ein - und das ist gut so!).

Welches Schuhsohle (A oder B) nutzt sich weniger ab?

Junge	1	2	3	4	5	6	7	8	9	10
A	13.2	8.2	10.9	14.3	10.7	6.6	9.5	10.8	8.8	13.3
B	14.2	8.8	11.2	14.2	11.8	6.4	9.8	11.3	9.3	13.6

Eine unpassende Analyse wäre es, hier auf die Paarung zu vergessen.

Dann geht auch die Variabilität zwischen den Individuen
in die Schätzung der Varianz ein,
was den Wert von f ungebührlich groß
und den Wert der t-Statistik klein macht.
Der t-Test für ungepaarte Stichproben ergibt einen
p-Wert von 0.72 - damit kann man nichts anfangen.

Beispiel 5

Wie untypisch ist die Lage der Ränge?

Der Wilcoxon-Test.

Wie eben zuvor geht es um einen Test der Hypothese,
dass zwei Stichproben
aus derselben Verteilung (auf $\mathbb{R}$) kommen,
gegen die Alternative, dass sich die beiden Verteilungen
durch eine Verschiebung unterscheiden.

Die folgende Idee kommt ganz

ohne spezielle Verteilungsannahme aus:

Man ordnet die $n_x + n_y$ Werte der Größe nach

und ersetzt sie durch ihre Ränge $R(x_i)$, $R(y_j)$.

(der kleinste Wert bekommt den Rang 1, der zweitkleinste den Rang 2,...).

Dann beobachtet man die *Rangsumme* $w := \sum_{i=1}^{n_x} R(x_i)$

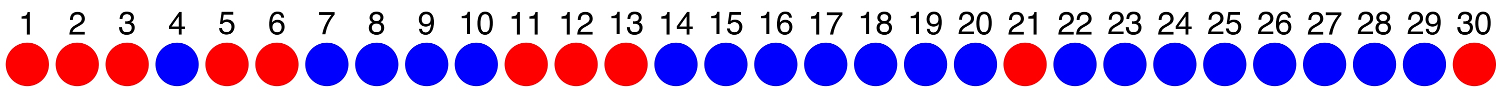
und fragt: Wie wahrscheinlich ist eine

mindestens so “randständige” Rangsumme

bei rein zufälliger Auswahl von n_x Elementen

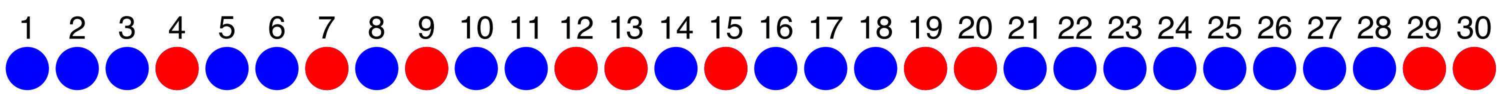
aus der Menge $\{1, \ldots, n_x + n_y\}$?

Die Raenge der x_i und der y_j



Rangsumme der x_i = 104

Eine zufaellige Permutation



Rangsumme der x_i in der Permutation = 158

Die “beobachtete” Rangsumme war 104.

Die minimale mögliche Rangsumme

einer “roten Teilstichprobe” ist $1 + \ldots + 10 = 55$.

Ihre maximale mögliche Rangsumme ist

$21 + \ldots + 30 = 255$.

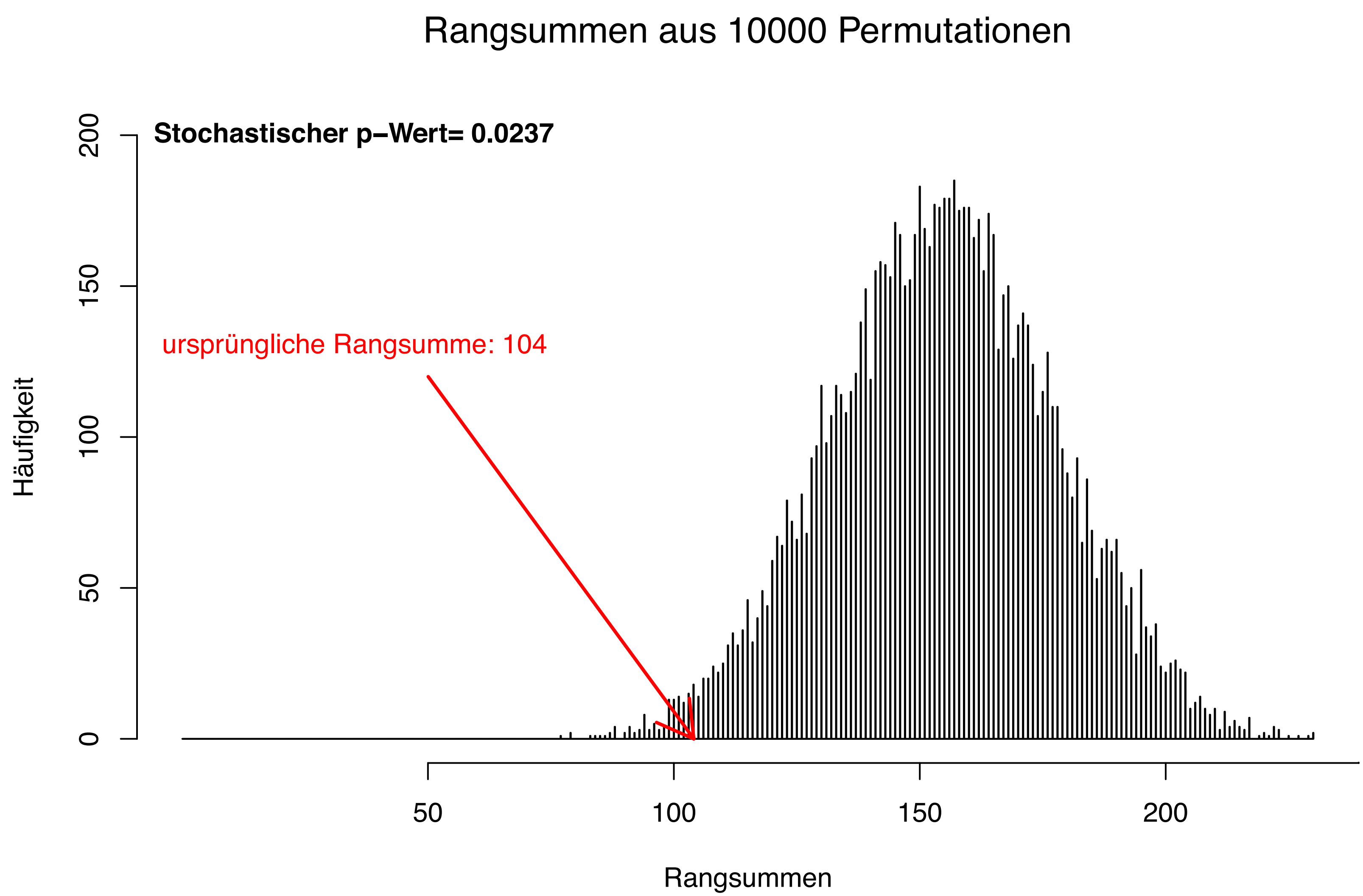
Wir ziehen 10000 mal eine Stichprobe der Größe 10 (aus 30)

und notieren deren Rangsumme.

Der *stochastische p-Wert* ist die relative Häufigkeit

der Ergebnisse, für die sich eine Rangsumme

≤ 104 oder $\geq 255 - (104 - 55)$ ergibt.


Rangsummen aus 10000 Permutationen
Stochastischer p−Wert= 0.0237
ursprüngliche Rangsumme: 104
Häufigkeit
0
50
100
150
200
50
100
150
200
Rangsummen